

Automatic Estimation of Flux Distributions of Astrophysical Source Populations

Raymond K. W. Wong* Paul Baines* Alexander Aue* Thomas C. M. Lee*
Vinay L. Kashyap[†]

May 4, 2013

Abstract

In astrophysics a common goal is to infer the flux distribution of populations of scientifically interesting objects such as pulsars or supernovae. In practice, inference for the flux distribution is often conducted using the cumulative distribution of the number of sources detected at a given sensitivity. The resulting “ $\log(N > S) - \log(S)$ ” relationship can be used to compare and evaluate theoretical models for source populations and their evolution. Under restrictive assumptions the relationship should be linear. In practice, however, when simple theoretical models fail, it is common for astrophysicists to use pre-specified piecewise linear models. This paper proposes a methodology for estimating both the number and locations of “breakpoints” in astrophysical source populations that extends beyond existing work in this field.

An important component of the proposed methodology is a new Interwoven EM Algorithm that computes parameter estimates. It is shown that in simple settings such estimates are asymptotically consistent despite the complex nature of the parameter space. Through simulation studies it is demonstrated that the proposed methodology is capable of accurately detecting structural breaks in a variety of parameter configurations. This paper concludes with an application of our methodology to the *Chandra* Deep Field North (CDFN) dataset.

Keywords: Broken power law, CDFN X-ray survey, Interwoven EM algorithm, Likelihood computations, $\log N - \log S$, Pareto distribution

1 Introduction

The relationship between the number of sources and the threshold at which they can be detected is an important tool in astrophysics for describing and investigating the properties of various types

*Department Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA. Email: [rkwwong,pdbaines,aaue,tcmlee]@ucdavis.edu.

[†]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA. Email: kashyap@head.cfa.harvard.edu.

of source populations. Known as the $\log N - \log S$ relationship, the idea is to use the number of sources $N(> S)$ that can be detected at a given sensitivity level S , on the log-log scale, to describe the distribution of source fluxes. In simple settings and under restrictive assumptions a linear relationship between the log-flux and the log-survival function can be derived from first principles. Traditionally astrophysicists have therefore examined this relationship by characterizing the slope of the log of the empirical survival function as a function of the log-flux of the sources.

Examples of $\log N - \log S$ analyses include Guetta *et al.* (2005), who use the relationship for Gamma Ray Bursts (GRBs) to constrain the structure of GRB jets. By comparing the $\log N - \log S$ relationship for observed data to the predicted $\log N - \log S$ relationship under different physical models for GRB jets, the authors are able to uncover limitations in the physical models. The $\log N - \log S$ curves have also been used to constrain cosmological parameters using cluster number counts in different passbands; see, e.g., Mathiesen and Evrard (1998) and Kitayama *et al.* (1998). Other applications of $\log N - \log S$ modeling include the study of active galactic nuclei (AGNs). For example, Mateos *et al.* (2008) use the $\log N - \log S$ relationship over different X-ray bands to constrain the population characteristics of hard X-ray sources.

In the probability domain under independent sampling, the linear $\log N - \log S$ relationship corresponds to a Pareto distribution for the source fluxes, known to astrophysicists as a power-law model. Despite the unrealistic assumptions in the derivation, the linear $\log N - \log S$ relationship does have strong empirical support in a variety of contexts; e.g., Kenter and Murray (2003). In addition to its simplicity the power-law model also retains a high degree of interpretability, with the power-law exponent often of direct scientific interest. As a result, of this simplicity and interpretability, the power-law model forms the basis of most $\log N - \log S$ analyses despite its many practical limitations in the ability to fit more complex datasets.

To address the limitations of this simple model astrophysicists have also experimented with a variety of broken power-law models. This is particularly important for larger populations or populations of sources spread over a wide energy range. Mateos *et al.* (2008) illustrate this by using both a two- and three-piece broken power-law model to capture the structure of the $\log N - \log S$ distribution across a wide range of energies. The basic idea of broken power-law models is to relax the assumption that the log survival function is a linear function of the log flux, and to instead assume a piecewise linear function. This adds additional challenges in estimating the location of the breakpoint, and quantifying the need for the breakpoint model above the simpler single power-law model. While recognizing the need to have more flexible models for $\log N - \log S$ analyses, most

of the work in this area does not provide a coherent means to selecting the location and number of breakpoints.

In this paper we provide an automatic method for jointly inferring the number and location of breakpoints and the parameters of interest for the $\log N - \log S$ problem. Our method allows astrophysicists to reliably infer both the number and the location of breakpoints in the $\log N - \log S$ relationship in a statistically rigorous manner for the first time. This simultaneous fitting introduces new computational challenges, so our method utilizes a new extension of the EM algorithm, known as the Interwoven EM Algorithm (IEM) (Baines, 2010; Baines *et al.*, 2012). The IEM algorithm provides efficient and stable estimation of the model parameters across a wide range of parameter settings for a fixed number of breakpoints. To determine the number of breakpoints we then use an additional model selection procedure that employs the power posterior technique of Friel and Pettitt (2008) to accurately compute the log-likelihood of the candidate models.

The remainder of the paper is organized as follows. In Section 2 we introduce the necessary background and statistical formulation of the $\log N - \log S$ model. Section 3 provides details of our estimation procedure for a fixed number of breakpoints, with Section 4 outlining our model selection procedure to determine the number of breakpoints required. The performance of our method in terms of both parameter estimation and identification of the number of breakpoints is detailed in Section 5. An application to data from the *Chandra* Deep-Field North X-ray survey is provided in Section 6. Large-sample theory is developed in Section 7 and concluding remarks are offered in Section 8. Lastly technical details are deferred to the appendix.

2 Background and Problem Specification

Let $\mathbf{S} = (S_1, \dots, S_n)^T$ denote a vector of the fluxes (in units of $\text{ergs s}^{-1} \text{ cm}^{-2}$) of each of a population of n astrophysical sources. For example, we may be interested in the flux distribution of a selection of n X-ray pulsars located in a specified region of sky at a specified distance. The basic building block of our method is the power-law model:

$$N(> S) = \sum_{i=1}^n I_{\{S_i > S\}} \simeq \alpha S^{-\beta}, \quad S > \tau. \quad (1)$$

This specifies that the unnormalized survival function $N(> S)$ is approximately a power of the flux S . The power-law exponent, β , is the parameter of primary interest and provides domain specific

knowledge about the source populations. The lower threshold τ can either be fixed according to the desired sensitivity level, or estimated from the data. Equivalently, taking the logarithm of both sides, (1) assumes a linear relationship between $\log(N(> S))$ and $\log(S)$:

$$\log(N(> S)) \simeq \log(\alpha) - \beta \log(S), \quad S > \tau. \quad (2)$$

In a statistical context, the theoretical power-law assumption corresponds to assuming that the source fluxes follow a Pareto distribution:

$$S_i \stackrel{\text{iid}}{\sim} \text{Pareto}(\beta, \tau), \quad i = 1, \dots, n.$$

In practice, the linear $\log N - \log S$, or Pareto, assumption is not sufficient to describe the $\log N - \log S$ relationship for many real datasets. There are several ways to generalize (1), the most popular among astrophysicists being the broken power-law model as illustrated in Jordán *et al.* (2004) and Cappelluti *et al.* (2007). The starting point of the broken power-law is to replace (1) with a monotonically decreasing piecewise linear approximation. In the case of a two-piece model we assume:

$$\log(N(> S)) = \begin{cases} \log(\alpha_1) - \beta_1 \log(S), & \tau_1 < S \leq \tau_2 \\ \log(\alpha_2) - \beta_2 \log(S), & S > \tau_2, \end{cases} \quad (3)$$

where β_1 and β_2 are parameters of interest. Note that as a result of the continuity and normalization constraints on $\tau_1, \tau_2, \alpha_1, \alpha_2, \beta_1$ and β_2 there are a total of 4 free parameters in this expanded two-piece model. Applications of the broken power-law model in the astrophysics community typically use either fixed numbers and locations of the breakpoint(s) or selection via ad hoc procedures (Trudolyubov *et al.*, 2002). The contribution of this paper is the proposal of an automatic procedure for selecting the number and estimating the locations of the breakpoints jointly with the parameters of interest.

2.1 Hierarchical modeling of the $\log N - \log S$ relationship

We now describe the connection between the broken power-law model introduced in (3) and the observed data. In practice the flux of each source, S_i , is not observed directly. Instead, we observe a Poisson-distributed photon count whose intensity is a known function of the parameter S_i . Let Y_1, Y_2, \dots, Y_n denote the source counts, then we assume the following hierarchical model. For

$i = 1, \dots, n$,

$$Y_i | S_1, \dots, S_n \stackrel{\text{indep.}}{\sim} \text{Poisson}(A_i S_i + b_i) \quad \text{and} \\ S_i \stackrel{\text{iid}}{\sim} \text{Pareto}_B(\boldsymbol{\beta}, \boldsymbol{\tau}),$$

where A_i 's and b_i 's are known constants (see below), $\boldsymbol{\beta} = (\beta_1, \dots, \beta_B) > \mathbf{0}$, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_B)$ such that $\tau_B > \dots > \tau_1 > 0$, and $\text{Pareto}_B(\boldsymbol{\beta}, \boldsymbol{\tau})$ represents a B -piece Pareto distribution with survival distribution

$$S_B(x) = \begin{cases} 1, & x < \tau_1 \\ \left(\frac{\tau_1}{x}\right)^{\beta_1}, & \tau_1 \leq x < \tau_2 \\ \left(\frac{\tau_1}{\tau_2}\right)^{\beta_1} \left(\frac{\tau_2}{x}\right)^{\beta_2}, & \tau_2 \leq x < \tau_3 \\ \vdots & \\ \left\{ \prod_{j=1}^{B-1} \left(\frac{\tau_j}{\tau_{j+1}}\right)^{\beta_j} \right\} \left(\frac{\tau_B}{x}\right)^{\beta_B}, & x \geq \tau_B \end{cases}$$

and thus its distribution function $F_B(\cdot) = 1 - S_B(\cdot)$. Note that the B -piece Pareto distribution corresponds to the broken power-law. The probability density f_B can be easily found by differentiation. When $B = 1$, the B -Pareto distribution reduces to a Pareto distribution with probability density function

$$f_1(x; \beta, \tau) = \begin{cases} \frac{\beta \tau^\beta}{x^{\beta+1}}, & x \geq \tau. \\ 0, & x < \tau. \end{cases}$$

In the above A_i 's, sometimes known as effective areas, represent sensitivities of the detector, while b_i 's represent background intensities. With the above model the goal is then to estimate B and, at the same time, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. At first sight, this seems to be a straightforward statistical problem: for a fixed B the maximum likelihood idea can be adopted to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, while the issue of choosing B can be viewed as a model selection problem and thus traditional ideas such as AIC and BIC can be used. However, as to be seen below, practical implementation of these ideas poses serious computational challenges that cannot be easily solved.

3 Maximum Likelihood Estimation When B Is Known

In this section we provide details of how to obtain maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ for a fixed number of breakpoints B in the $\log N - \log S$ model. Defining $\beta_0 = 0$, $\tau_0 = \tau_1$ and $\tau_{B+1} = \infty$,

the likelihood is

$$L(\boldsymbol{\beta}, \boldsymbol{\tau}; Y_1, \dots, Y_n) = \prod_{i=1}^n \left\{ \int_{\tau_1}^{\infty} \frac{e^{-(A_i s + b_i)} (A_i s + b_i)^{Y_i}}{Y_i!} f_B(s; \boldsymbol{\beta}, \boldsymbol{\tau}) ds \right\}.$$

Note that the likelihood involves some numerically unstable integrals that do not have a closed form solution, and hence a direct maximization is extremely difficult. To further appreciate this difficulty, consider the case when there is no background contamination ($b_i = 0$), for which the above likelihood degenerates to

$$\prod_{i=1}^n \left[\sum_{j=1}^B \left(\frac{\tau_{j-1}}{\tau_j} \right)^{\beta_{j-1}} \frac{\beta_j (A_i \tau_j)^{\beta_j}}{Y_i!} \{ \Gamma(Y_i - \beta_j, A_i \tau_j) - \Gamma(Y_i - \beta_j, A_i \tau_{j+1}) \} \right].$$

Here, $\Gamma(a, x) = \int_x^{\infty} t^{a-1} e^{-t} dt$ is the gamma function which is numerically unstable, particularly when the first argument is large. Together with the inner summation in the above expression, these issues make a direct maximization of the (log-)likelihood difficult even when there is no background contamination. To address these issues we propose an EM-algorithm (Dempster *et al.*, 1977) to find the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ for the general case of $b_i \geq 0$.

3.1 EM with Sufficiency Augmentation Scheme

The EM algorithm (Dempster *et al.*, 1977) has long been popular for its monotone convergence and resulting stability, and is therefore well-suited to our context. As always, the EM algorithm must be formulated in terms of “missing data” or auxiliary variables, that must be integrated out to obtain the observed data log-likelihood. For the current problem, since we are interested only in inference for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, marginalizing over the uncertainty in the individual fluxes, it is natural to treat $\mathbf{S} = (S_1, \dots, S_n)^T$ as the missing data. Since \mathbf{S} is a sufficient statistic for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})^T$, we call this the sufficient augmentation (SA) scheme in the terminology of Yu and Meng (2011).

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. The complete data log-likelihood of (\mathbf{Y}, \mathbf{S}) is

$$\log p(\mathbf{Y}, \mathbf{S}; \boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{i=1}^n \log g(Y_i; A_i S_i + b_i) + \sum_{i=1}^n \log f_B(S_i; \boldsymbol{\beta}, \boldsymbol{\tau}),$$

where $g(x; \mu)$ is the probability mass function of a Poisson distribution with mean μ . In the E-Step

of the algorithm we compute the conditional expectation

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= \mathbb{E} \left\{ \log p(\mathbf{Y}, \mathbf{S}; \boldsymbol{\theta}) | \mathbf{Y}; \boldsymbol{\theta}^{(k)} \right\} \\ &= \sum_{i=1}^n \mathbb{E} \left\{ \log g(Y_i; A_i S_i + b_i) | Y_i; \boldsymbol{\theta}^{(k)} \right\} + \sum_{i=1}^n \mathbb{E} \left\{ \log f_B(S_i; \boldsymbol{\theta}) | Y_i; \boldsymbol{\theta}^{(k)} \right\}, \end{aligned} \quad (4)$$

where $\boldsymbol{\theta}^{(k)}$ denotes the estimate of $\boldsymbol{\theta}$ at the k -th iteration. The M-step of the algorithm must then maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$. Since the first term of (4) does not depend on $\boldsymbol{\theta}$, it can be ignored in our maximization. For the second term, as it does not admit a closed form expression, a Monte Carlo method is used to approximate it. The basic idea is to estimate it by the mean of a suitable Monte Carlo sample of the S_i 's as described in Algorithm 1.

Without the first term in (4), the maximization of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ is equivalent to finding the MLE of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})^T$ from an iid sample $\mathbf{X} = (X_1, \dots, X_m)$ from the $\text{Pareto}_B(\boldsymbol{\beta}, \boldsymbol{\tau})$ distribution. The log-likelihood of \mathbf{X} is

$$l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{j=1}^B \beta_j (n_j \log \tau_j - n_{j+1} \log \tau_{j+1}) + \sum_{j=1}^B m_j \log \beta_j - \sum_{j=1}^B \beta_j \sum_{i \in A_j} \log X_i - \sum_{i=1}^m \log X_i,$$

where $n_j = \text{card}\{i: X_i \geq \tau_j\}$, $n_{B+1} = 0$, $m_j = n_{j+1} - n_j$, $\tau_{B+1} = \infty$, $n_{B+1} \log \tau_{B+1}$ is defined to be 0, and $A_j = \{i: \tau_j \leq X_i < \tau_{j+1}\}$. Note that the n_j 's and m_j 's are functions of $\boldsymbol{\tau}$. For any fixed $\boldsymbol{\tau}$, straightforward algebra shows that $l(\boldsymbol{\theta}; \mathbf{X})$ is maximized when β_j is set to

$$\beta_j(\boldsymbol{\tau}) = m_j(\boldsymbol{\tau}) \left(\sum_{i \in A_j} \log X_i + n_{j+1}(\boldsymbol{\tau}) \log \tau_{j+1} - n_j(\boldsymbol{\tau}) \log \tau_j \right)^{-1}, \quad j = 1, \dots, B. \quad (5)$$

By substituting the above expression, $l(\boldsymbol{\theta}; \mathbf{X})$ becomes

$$l(\boldsymbol{\theta}; \mathbf{X}) = -m - \sum_{i=1}^m \log X_i + \sum_{j=1}^B m_j(\boldsymbol{\tau}) \log \beta_j(\boldsymbol{\tau}). \quad (6)$$

Therefore, to obtain the MLE for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})^T$ from \mathbf{X} , one can first maximize $l(\boldsymbol{\theta}; \mathbf{X})$ in (6) with respect to $\boldsymbol{\tau}$, and then plug in the corresponding maximizer $\hat{\boldsymbol{\tau}}$ (i.e., the MLE of $\boldsymbol{\tau}$) into (5) to obtain the MLE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

The MLE of τ_1 is $\hat{\tau}_1 = \min(X_1, \dots, X_m)$, while unfortunately the MLEs for τ_2, \dots, τ_B do not admit closed-form expressions. Further, (6) is not a continuous function in $\boldsymbol{\tau}$ and therefore

traditional optimization methods that require function derivatives (e.g., Newton-like methods) cannot be applied here. We have experimented with various optimization algorithms and found that the Nelder-Mead algorithm works well for this problem. The major steps of the EM algorithm in the SA scheme (SAEM) for finding the MLEs of $\boldsymbol{\theta}$ are given in Algorithm 1. In practice, the SAEM algorithm often converges very slowly. Section 3.4 below provides some illustrative numerical examples.

Algorithm 1 SAEM: EM with the Sufficient Augmentation Scheme (SAEM)

1. Choose a starting value $\boldsymbol{\theta}^{(0)}$ and set $k = 0$.
2. Generate $\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(N_{\text{sim}})}$ from $p(\mathbf{S}|\mathbf{Y}; \boldsymbol{\theta}^{(k)})$ using the following Metropolis-Hastings algorithm. For each simulation of \mathbf{S} , we sample the elements of \mathbf{S} one at a time. Suppose $\mathbf{S} = (S_1, \dots, S_n)$ is the current draw. Denote $\mathbf{S}^* = (S_1, \dots, S_{j-1}, S_j^*, S_{j+1}, \dots, S_n)$, where S_j^* is drawn from $\text{Pareto}_B(\boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)})$. We accept this \mathbf{S}^* as new value with probability $a_j(\mathbf{S}, \mathbf{S}^*)$; otherwise, we retain \mathbf{S} . The acceptance probability is given by

$$a_j(\mathbf{S}, \mathbf{S}^*) = \min \left\{ 1, \frac{g(Y_j; A_j S_j^* + b_j)}{g(Y_j; A_j S_j + b_j)} \right\}.$$

3. Find the maximizer $\tilde{\boldsymbol{\theta}}$ of the Monte Carlo estimate of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$. This is equivalent to computing

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \frac{1}{N_{\text{sim}} - N_{\text{burn}}} \sum_{s=N_{\text{burn}}+1}^{N_{\text{sim}}} \sum_{i=1}^n \log f_B(S_i^{(s)}; \boldsymbol{\theta}),$$

where N_{burn} is the number of burn-in. As discussed above, $\tilde{\boldsymbol{\theta}}$ can be obtained by the following steps:

- (a) set $\tilde{\tau}_1 = \min\{S_i^{(s)} : i = 1, \dots, n, s = N_{\text{burn}} + 1, \dots, N_{\text{sim}}\}$,
 - (b) obtain $\tilde{\tau}_2, \dots, \tilde{\tau}_B$ as the maximizer of $\sum_{j=1}^B m_j(\boldsymbol{\tau}^*) \log \beta_j(\boldsymbol{\tau}^*)$, where $\boldsymbol{\tau}^* = (\tilde{\tau}_1, \tau_2, \dots, \tau_B)$, using the Nelder-Mead algorithm, and
 - (c) set $\tilde{\beta}_j = \beta_j(\tilde{\boldsymbol{\tau}})$ using (5), for $j = 1, \dots, B$.
4. Set $\boldsymbol{\theta}^{(k+1)} = \tilde{\boldsymbol{\theta}}$.
 5. Repeat Steps 2 to 4 until convergence.
-

3.2 EM with Ancillary Augmentation Scheme (AAEM)

Given the slow convergence of the SAEM algorithm, we seek faster alternatives. This subsection proposes an alternative EM algorithm that is based on an ancillary augmentation (AA) scheme,

called the AAEM algorithm. For a discussion of augmentation schemes and their use in EM, see Baines *et al.* (2012). The basis of our AAEM is to re-express our model using auxiliary variables $U_i = F_B(S_i; \boldsymbol{\theta})$:

$$Y_i | U_1, \dots, U_n \stackrel{\text{indep.}}{\sim} \text{Poisson}(A_i F_B^{-1}(U_i; \boldsymbol{\theta}) + b_i) \quad \text{and} \\ U_i \stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1),$$

for $i = 1, \dots, n$. Here $\mathbf{U} = (U_1, \dots, U_n)$ is treated as the missing data, and preserves the observed data log-likelihood. In the E-Step we then calculate the conditional expectation

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k)}) = \sum_{i=1}^n \mathbb{E} \left\{ \log g(Y_i; A_i F_B^{-1}(U_i; \boldsymbol{\theta}) + b_i) | Y_i; \boldsymbol{\theta}^{(k)} \right\}. \quad (7)$$

This conditional expectation can be approximated and maximized in a similar manner as for the $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k)})$ in the SAEM algorithm. The resulting AAEM algorithm is summarized in Algorithm 2. Section 3.4 provides some empirical comparisons between the AAEM and SAEM algorithms. As may be expected, there are some situations where the AAEM algorithm converges faster, while there are other situations where the SAEM algorithm converges faster.

3.3 Interwoven EM (IEM)

In practice, choosing the most efficient algorithm between the SAEM and AAEM requires knowledge of the unknown parameter values and the theoretical convergence rates, both of which are not available in most contexts. Therefore, it would instead be desirable if one could combine the “best parts” of SAEM and AAEM rather than select one of them. One simple way to combine the two algorithms is to use the so-called alternating EM (AEM) algorithm. The AEM algorithm proceeds by using SAEM for the first iteration, then uses AAEM for the second iteration, followed by SAEM for the third, and so on. While this procedure tends to “average” the performance of the two algorithms, a more sophisticated way to combine them is to use the Interwoven EM (IEM) algorithm of Baines *et al.* (2012). Theoretical and empirical results show that IEM typically achieves sizeable performance gains over the component EM algorithms. The key to the boosted performance of IEM is that it utilizes the joint structure of the two augmentation schemes through a special “IE-Step”. In contrast, AEM simply performs sequential updates using each augmentation scheme that make no use of this joint information. The theory of the IEM algorithm in Baines *et al.*

Algorithm 2 AAEM: EM with Ancillary Augmentation Scheme

1. Choose a starting value $\boldsymbol{\theta}^{(0)}$ and set $k = 0$.
2. Generate $\mathbf{U}^{(1)}, \dots, \mathbf{U}^{(N_{\text{sim}})}$ from $p(\mathbf{U}|\mathbf{Y}; \boldsymbol{\theta}^{(k)})$ using the Metropolis-Hastings algorithm. For each simulation of \mathbf{U} , we sample the element of \mathbf{U} one by one. Let $\mathbf{U} = (U_1, \dots, U_n)$ be the previous draw. If we denote $\mathbf{U}^* = (U_1, \dots, U_{j-1}, U_j^*, U_{j+1}, \dots, U_n)$, where U_j^* is drawn from $\text{Uniform}(0, 1)$. We accept this \mathbf{U}^* as new value with probability $b_j(\mathbf{U}, \mathbf{U}^*)$; otherwise, we retain \mathbf{U} . The acceptance probability is given by

$$b_j(\mathbf{U}, \mathbf{U}^*) = \min \left\{ 1, \frac{g(Y_j; A_j F_B^{-1}(U_j^*; \boldsymbol{\theta}^{(k)}) + b_j)}{g(Y_j; F_B^{-1}(U_j; \boldsymbol{\theta}^{(k)}) + b_j)} \right\}.$$

3. Find the maximizer $\tilde{\boldsymbol{\theta}}$ of the following Monte Carlo estimate of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$:

$$\frac{1}{N_{\text{sim}} - N_{\text{burn}}} \sum_{s=N_{\text{burn}}+1}^{N_{\text{sim}}} \sum_{i=1}^n \log g(Y_i; A_i F_B^{-1}(U_i^{(s)}; \boldsymbol{\theta}) + b_i).$$

The maximization can be done for example with the Nelder-Mead algorithm.

4. Set $\boldsymbol{\theta}^{(k+1)} = \tilde{\boldsymbol{\theta}}$.
 5. Repeat Steps 2 to 4 until convergence.
-

(2012) shows that the rate of convergence of IEM is dependent on the “correlation” between the two component augmentation schemes. Since the SA and AA schemes typically have low correlation, here we interweave these two schemes to produce an IEM algorithm for estimating the parameters of flux distributions.

The IEM algorithm for our $\log N - \log S$ model is given in Algorithm 3. The algorithm requires very minimal computation in addition to the component SAEM and AAEM algorithms so is comparable in real-time per-iteration speed. Lastly we note that there is some freedom in how to combine the IEM algorithm with MC methods. Specifically, there are variations in how one may choose to implement Step 3. One may want to sample U again instead of using the previous samples in Step 2. In both cases, one obtains a sample from $\mathbf{U}|\mathbf{Y}, \boldsymbol{\theta}^{(k+0.5)}$ and achieves the goal. From our practical experience, we found that there is very little difference between the performances of these two approaches. Thus, we choose to use the one which is least computationally expensive.

Algorithm 3 IEM: Interwoven EM

1. Choose a starting value $\theta^{(0)}$ and set $k = 0$.
 2. Execute Steps 2 and 3 of the SAEM algorithm. Set $\theta^{(k+0.5)} = \tilde{\theta}$.
 3. Execute Step 3 of the AAEM algorithm, with $\mathbf{U}^{(l)}$ generated as: $U_j^{(l)} = F_B(S_j^{(l)}; \theta^{(k+0.5)})$, for $j = 1, \dots, n$ and $l = N_{\text{burn}} + 1, \dots, N_{\text{sim}}$. Set $\theta^{(k+1)} = \tilde{\theta}$.
 4. If convergence is achieved or k attains N_{limit} , then declare $\theta^{(k+1)}$ to be MLE; otherwise set $k = k + 1$ and return to Step 2.
-

3.4 An Empirical Comparison Amongst Different EM Algorithms

In this subsection we empirically compare the convergence speeds of SAEM, AAEM, AEM and IEM by applying them to two simulated data sets. These two data sets were simulated from a model with $B = 1$ and no background contamination counts. This model is somewhat simple but the advantage is that the likelihood function simplifies considerably, and the corresponding maximum likelihood estimates can be reliably obtained with non-EM methods. With these maximum likelihood estimates the maximized log-likelihood value can be calculated and used for baseline comparisons.

In Figure 1(a), for the first simulated data set, we plot the negative log-likelihood values of the SAEM, AAEM, AEM and IEM estimates evaluated at different iterations. One can see the slow convergence speeds of SAEM and AAEM, with SAEM being the slower. Also, both AEM and IEM converged relatively fast, with IEM being the faster. When comparing to AEM, IEM utilizes the relationship between SAEM and AAEM at each step, which leads to the superiority of IEM over AEM.

We repeat the same plot in Figure 1(b) for the second simulated data set. This time the relative speeds of SAEM and AAEM switched; i.e., SAEM converged faster. This illustrates the fact that none of SAEM and AAEM is superior to the other. The relative position of AEM and IEM remain the same.

Overall from these two plots one can see that the IEM algorithm is the most efficient and robust. Also, when comparing to AEM, it is computationally faster due to the skipping of an extra sampling step. Similar performance was observed across a wide range of simulation settings. Therefore we recommend using the IEM algorithm to compute the maximum likelihood estimates when B is known.

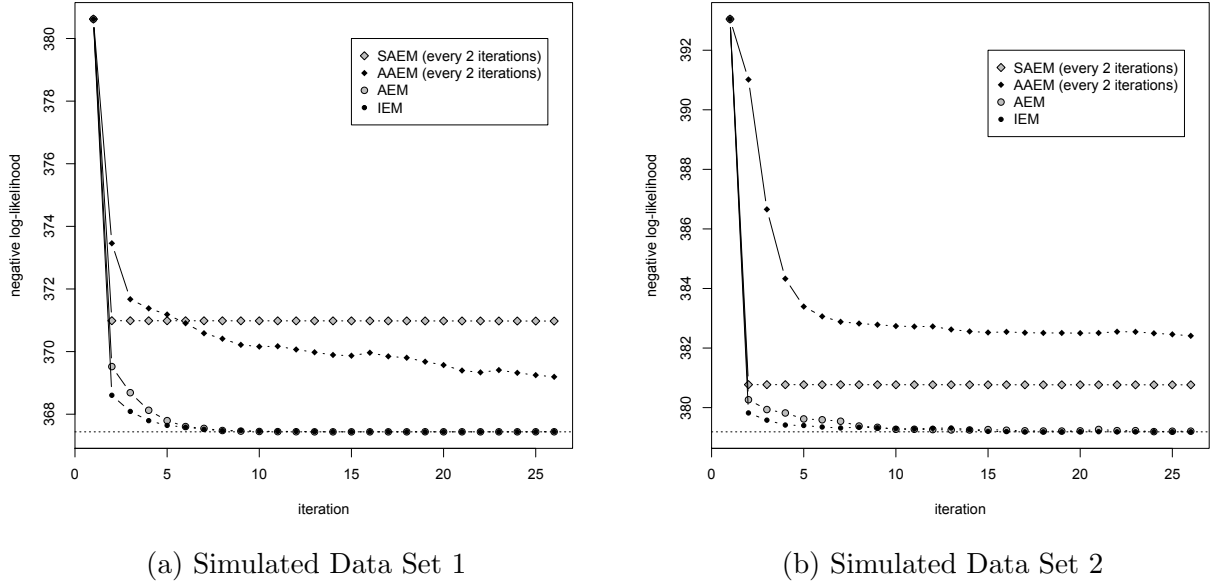


Figure 1: Plots of negative log-likelihood values for different EM algorithms. In each plot the horizontal dashed line indicates the negative log-likelihood evaluated at the maximum likelihood estimates.

4 Automated Choice of B

This section addresses the important problem of selecting the number of “pieces”, B , in the broken-Pareto model. Since this problem can be seen as a model selection problem, one can adopt well studied methods such as AIC and BIC to solve it. To proceed we first note that when $B = 1$, the number of free parameters in the model is $2B$. With AIC, the best B is chosen as

$$\hat{B}_{\text{AIC}} = \operatorname{argmax}_B \text{AIC}(B) = \operatorname{argmax}_B \left\{ -2 \log L(\hat{\beta}, \hat{\tau}; Y_1, \dots, Y_n) + 4B \right\},$$

while for BIC B is chosen as the minimizer of

$$\hat{B}_{\text{BIC}} = \operatorname{argmax}_B \text{BIC}(B) = \operatorname{argmax}_B \left\{ -2 \log L(\hat{\beta}, \hat{\tau}; Y_1, \dots, Y_n) + 2B \log n \right\}.$$

Despite the straightforward definitions, in practice the numerical instability of the likelihood function makes computation of $\text{AIC}(B)$ and $\text{BIC}(B)$ very challenging. To address this problem we adopt the so-called power posterior method proposed by Friel and Pettitt (2008) to approximate the log-likelihood directly.

In our context, the power posterior is defined as

$$p_t(\mathbf{S}|\mathbf{Y}; \boldsymbol{\theta}) \propto p(\mathbf{Y}|\mathbf{S})^t p(\mathbf{S}; \boldsymbol{\theta}) \quad \text{for } 0 \leq t \leq 1.$$

In addition, define

$$z(\mathbf{Y}|t) = \int_{\mathbb{R}^n} p(\mathbf{Y}|\mathbf{s})^t p(\mathbf{s}; \boldsymbol{\theta}) d\mathbf{s},$$

and, for simplicity, write the likelihood as $p(\mathbf{Y}) = L(\boldsymbol{\beta}, \boldsymbol{\tau}; Y_1, \dots, Y_n)$. The following equality is crucial to this method:

$$\log\{p(\mathbf{Y})\} = \log \left\{ \frac{z(\mathbf{Y}|t=1)}{z(\mathbf{Y}|t=0)} \right\} = \int_0^1 \mathbb{E} [\log\{p(\mathbf{Y}|\mathbf{S})\} | \mathbf{Y}; \boldsymbol{\theta}, t] dt,$$

where the last expectation (inside the integral) is taken with respect to the power posterior $p_t(\mathbf{S}|\mathbf{Y}; \boldsymbol{\theta})$. The idea is as follows. First, for any given t , Monte Carlo methods can be applied to sample from the power posterior and approximate the expectation. Once a sufficient number of these expectations (corresponding to different values of t) are calculated, numerical methods can be used to approximate the integral, which is the same as the log-likelihood. Since this method approximates the log-likelihood directly (i.e., without the computation of the likelihood), it is numerically quite stable. The detailed algorithm is presented as Algorithm 4.

The above algorithm provides a reliable method for approximating the log-likelihood for a given value of $\boldsymbol{\theta}$. Then one natural question to ask is, can we not simply obtain the MLE of $\boldsymbol{\theta}$ by directly maximizing this log-likelihood approximation via, say, Newton's method? The answer, in principle, is yes, but the IEM algorithm is still preferred mainly because the estimates from IEM are generally more stable and reliable. Moreover, the power posterior approximation to the log-likelihood is computationally intensive if one wants to obtain an accurate estimate. For these reasons, we only use this power posterior approximation to estimate the log-likelihood evaluated at the MLE obtained by the IEM algorithm.

5 Simulation Experiments

Numerical experiments were conducted to evaluate the practical performance of the proposed methodology. Four experimental settings were considered:

1. $B = 1$, $\boldsymbol{\tau} = 5 \times 10^{-17}$, $\boldsymbol{\beta} = 1$ and $n = 100$,

Algorithm 4 Power Posterior Method for Log-Likelihood Calculation

1. Choose a starting value $\mathbf{S}^{(0)}$ and set $k = 0$.
2. Set $t = (k/N_{\text{grid}})^c$, where c controls the density of the grid values of t . It is typically set to 3 or 5 (see Friel and Pettitt, 2008).
3. Generate $\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(N_{\text{sim}})}$ from $p_t(\mathbf{S}|\mathbf{Y}; \boldsymbol{\theta})$ using the Metropolis-Hastings algorithm described in Step 2 of the SAEM algorithm. Note that the acceptance probability becomes

$$a_j(\mathbf{S}, \mathbf{S}^*) = \min \left\{ 1, \left\{ \frac{g(Y_j; A_j S_j^* + b_j)}{g(Y_j; A_j S_j + b_j)} \right\}^t \right\}.$$

4. Estimate $\mathbb{E}[\log\{p(\mathbf{Y}|\mathbf{S})\}|\mathbf{Y}; \boldsymbol{\theta}, t]$ with

$$\hat{l}_t = \frac{1}{N_{\text{sim}} - N_{\text{burn}}} \sum_{s=N_{\text{burn}}+1}^{N_{\text{sim}}} \log p(\mathbf{Y}|\mathbf{S}^{(s)}; \boldsymbol{\theta}).$$

5. If $k < N_{\text{grid}}$, set $k = k+1$, $\mathbf{S}^{(0)} = \sum_{s=N_{\text{burn}}+1}^{N_{\text{sim}}} \mathbf{S}^{(s)} / (N_{\text{sim}} - N_{\text{burn}})$, and go to Step 2. Otherwise go to the next step.
 6. Given the \hat{l}_t 's, the log-likelihood $\log\{p(\mathbf{Y})\}$ can be approximated via any reliable numerical integration method.
-

2. $B = 2$, $\boldsymbol{\tau} = (1 \times 10^{-17}, 5 \times 10^{-17})^T$, $\boldsymbol{\beta} = (0.5, 3)^T$ and $n = 200$,
3. $B = 2$, $\boldsymbol{\tau} = (1 \times 10^{-17}, 5 \times 10^{-17})^T$, $\boldsymbol{\beta} = (0.5, 1.5)^T$ and $n = 200$,
4. $B = 3$, $\boldsymbol{\tau} = (1 \times 10^{-17}, 8 \times 10^{-17}, 1.8 \times 10^{-16})^T$, $\boldsymbol{\beta} = (0.3, 1, 3)^T$ and $n = 500$.

The parameter values of these settings were chosen to mimic the typical behavior of the real data. The effective areas and the expected background counts are set to $A_i = 10^{19}$ and $b_i = 10$ respectively for all i .

Two hundred data sets were generated for each experimental setting. For each generated data set, both AIC and BIC were applied to choose the value of B , and model parameters were estimated by the IEM algorithm. The selected values of B are summarized in Table 1. One can see that BIC works substantially better than AIC for selecting B , and while BIC occasionally overestimates B , there is a clear tendency for AIC to consistently overestimate B .

Other crucial factors that determine the ability of our method to detect structural breaks in the population distribution include: (i) the sample size, (ii) the separation between breakpoints, and, (iii) the magnitude of the difference between the power-law slopes on adjacent segments. The

impact of the third factor can be seen by comparing simulation results from settings 2 and 3, where the misclassification rate is seen to increase as the slopes become closer. From additional simulations our experience suggests that in typical settings a sample size of 200 or more is needed to reliably detect a single breakpoint, with double this required to detect two breakpoints. In simulations, true breakpoints can be detected for smaller sample sizes, but at a lower rate that is more dependent on the noise properties of the specific simulation.

In addition to selecting the number of breakpoints, we also conducted a simulation to assess the quality of parameter estimation when using the IEM algorithm. For each experimental setting, we calculated the squared error $(\beta_1 - \hat{\beta}_1)^2$ of $\hat{\beta}_1$ for all those data sets where \hat{B} were correctly selected. We then computed the average of all these squared errors, denoted as $\text{mse}(\hat{\beta}_1)$, and calculated the relative mean squared error $\sqrt{\text{mse}(\hat{\beta}_1)}/\beta_1$. Similar relative mean squared errors for other estimates in $\hat{\beta}$ and $\hat{\tau}$ were obtained in a similar manner. These relative mean squared errors are given in Table 2. We note that all of these are of the order of 10^{-2} or 10^{-1} .

| Experimental Setting | Model Selection Method | \hat{B} | | | |
|----------------------|------------------------|-----------|-----|-----|----|
| | | 1 | 2 | 3 | 4 |
| 1 | AIC | 94 | 53 | 35 | 18 |
| | BIC | 164 | 33 | 3 | 0 |
| 2 | AIC | 0 | 135 | 45 | 20 |
| | BIC | 0 | 198 | 2 | 0 |
| 3 | AIC | 0 | 110 | 71 | 19 |
| | BIC | 0 | 177 | 23 | 0 |
| 4 | AIC | 0 | 0 | 138 | 62 |
| | BIC | 0 | 0 | 194 | 6 |

Table 1: The number of pieces \hat{B} selected by AIC and BIC.

6 Application: *Chandra* Deep Field North X-Ray Data

We now apply our method to data from the *Chandra* Deep Field North (CDFN) X-ray survey. Our dataset comprises a total of 225 sources with an off-axis angle of 8 arcmins or less and counts ranging from 5 to 8655. The full CDFN dataset is comprised of multiple observations at many different aimpoints, however we here consider only a subset where the aimpoints are close to each other to avoid additional modeling considerations. Since off-axis angle measures the radial distance

| Experimental Setting | Model Selection Method | $\hat{\tau}$ | | | $\hat{\beta}$ | | |
|----------------------|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 1 | AIC | 5.14×10^{-2} | - | - | 1.11×10^{-1} | - | - |
| | BIC | 4.91×10^{-2} | - | - | 1.06×10^{-1} | - | - |
| 2 | AIC | 3.33×10^{-2} | 2.55×10^{-2} | - | 9.81×10^{-2} | 1.13×10^{-1} | - |
| | BIC | 3.52×10^{-2} | 2.60×10^{-2} | - | 9.17×10^{-2} | 1.08×10^{-1} | - |
| 3 | AIC | 3.52×10^{-2} | 1.42×10^{-1} | - | 1.20×10^{-1} | 1.32×10^{-1} | - |
| | BIC | 3.57×10^{-2} | 1.29×10^{-1} | - | 1.11×10^{-1} | 1.35×10^{-1} | - |
| 4 | AIC | 2.71×10^{-2} | 3.26×10^{-2} | 5.04×10^{-2} | 7.08×10^{-2} | 9.91×10^{-2} | 1.23×10^{-1} |
| | BIC | 2.72×10^{-2} | 3.94×10^{-2} | 4.97×10^{-2} | 7.16×10^{-2} | 9.74×10^{-2} | 1.19×10^{-1} |

Table 2: The relative mean squared errors of $\hat{\beta}$ and $\hat{\tau}$, conditional on selection of the correct B .

of the source from the center of the detector, sources with large off-axis angles can be thought of as being “close to the edge of the image”. Sources appearing at large off-axis angles appear much larger and at lower resolution than those closer to the center of the detector. The source-specific scaling constant, effective area A_i , is used to account for variations in the expected number of photons as a function of source location and photon energy. However, at large off-axis angles additional complications such as “confusion” (two or more sources overlapping and appearing as one) and “incompleteness” (possible non-detections of fainter sources) must be considered. For the purposes of our analysis here, we include all sources with an off-axis angle < 8 arcmin to achieve a worst-case completeness of 80%. We also consider thresholding at < 6 and < 7 arcmins, with a full discussion of the sensitivity to this threshold considered in Section 6.1.

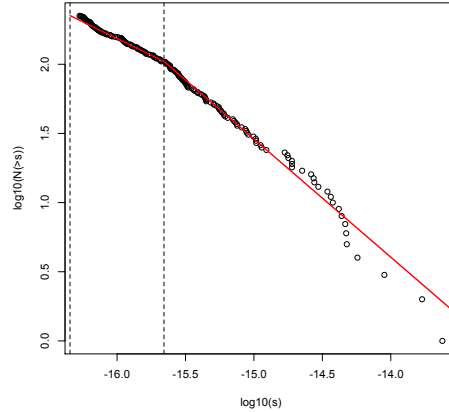


Figure 2: $\log N - \log S$ plot for the *Chandra* Deep Field North data with off-axis angle truncation at 8 arcmins. The vertical dotted lines are drawn at $\hat{\tau}_1$ and $\hat{\tau}_2$. The red lines correspond to the fitted broken-Pareto model with estimated slopes $\hat{\beta}_1$ and $\hat{\beta}_2$.

Applying our model selection procedure to the dataset with < 8 arcmins yields an estimate of $\hat{B} = 2$, with $\hat{B} = 1$ for the < 6 and < 7 arcmin subsets. As discussed in detail in Section 6.1, the consistency of the observations in the $6 - 8$ arcmin range suggests that the ability to detect the presence of a breakpoint is limited by the small sample sizes at < 6 and < 7 . Figure 2 shows the $\log N - \log S$ plot for the < 8 arcmin dataset, depicting the log (base 10) of the empirical survival count as a function of the log flux, using the imputed fluxes from the final E-step of our algorithm. While the plot ignores the uncertainty in the S_i 's, it remains the standard plot for the analysis of $\log N - \log S$ relationships. We note from the plot that the “break” is clearly visible around $\log_{10}(\tau_1) = -15.657$, with a change in slope from 0.48 to 0.85. Full parameter estimates and standard error estimates are provided in Table 3. Standard error estimates are obtained using a simple Bootstrap resampling procedure. Our analysis shows that a two-piece broken power-law model is preferred for this subset, with a breakpoint at a lower flux than shown in Moretti *et al.* (2003), and with the lower segment at a flatter slope. This differs from what would be expected if point sources are to make up all of the diffuse background (Hickox and Markevitch, 2007), suggesting that a significant proportion of the residual X-ray background is composed of diffuse emission; see also Mateos *et al.* (2008).

| Parameter | Estimate | SE |
|---------------------|----------|-------|
| β_1 | 0.483 | 0.060 |
| β_2 | 0.854 | 0.224 |
| $\log_{10}(\tau_1)$ | -16.344 | 0.030 |
| $\log_{10}(\tau_2)$ | -15.657 | 0.271 |

Table 3: Parameter estimates and standard errors for the *Chandra* Deep Field North (CDFN) dataset.

6.1 CDFN Source Selection

In this section we consider the sensitivity of our analysis to the chosen off-axis angle threshold. As discussed in Section 6, at higher off-axis angles there are additional complications such as incompleteness and confusion that must be built into any statistical analysis that are not covered by the method presented here. Let K denote the maximum off-axis angle; i.e., all sources with off-axis angle less than K are retained, all others are excluded from the analysis. The choice of $K = 8$ for our analysis in Section 6 is motivated by scientific considerations and an estimated completeness above 80% at $K = 8$. However, by varying the truncation point we obtain additional

| K | n | $\log_{10}(\hat{\tau})$ | | $\hat{\beta}$ | |
|-----|-----|---------------------------|---------------------------|-----------------|-----------------|
| | | $\log_{10}(\hat{\tau}_1)$ | $\log_{10}(\hat{\tau}_2)$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ |
| 4 | 77 | -16.364 | | 0.788 | |
| 5 | 112 | -16.353 | | 0.738 | |
| 6 | 152 | -16.329 | | 0.691 | |
| 7 | 192 | -16.373 | | 0.590 | |
| 8 | 225 | -16.343 | -15.668 | 0.482 | 0.850 |
| 9 | 257 | -16.352 | -15.732 | 0.449 | 0.850 |
| 10 | 287 | -16.378 | -15.696 | 0.450 | 0.792 |
| 11 | 298 | -16.389 | -15.702 | 0.456 | 0.793 |
| 12 | 303 | -16.403 | -15.677 | 0.454 | 0.802 |
| 13 | 304 | -16.429 | -15.843 | 0.412 | 0.743 |

Table 4: CDFN Results by varying off-axis truncation

insight into the sensitivity of our analysis to this decision, as well as to the statistical sensitivity to the sample size required for breakpoint detection. Table 4 shows the results of the analysis for differing values of K . As explained, results for $K > 9$ are likely to be untrustworthy, although they happen to be similar to those with $K = 8$. On the other extreme, if we truncate at $K = 4$ or $K = 5$ we unnecessarily discard a large number of sources.

We note that at $K = 7$ we are also no longer able to formally detect a break i.e., $\hat{B} = 1$. However, upon closer examination the BIC values for $B = 1$ and $B = 2$ when $K = 7$ are very similar (2186.79 vs. 2188.37), indicating that there is little to choose between the $B = 1$ and $B = 2$ models. With a few additional data points added at $K = 8$, our procedure then has enough power to detect the break at $K = 8$. It is worth noting that all additional data points with off-axis angle between 7 and 8 were manually screened, and are quantitatively very similar to those with $K < 7$. That is, the detection (or lack) of a breakpoint in this context appears to be primarily determined by the sample size of the dataset used. This is consistent with our results from the simulation study in Section 5, where a sample size of approximately 200 was required to reliably detect a break with similar parameter configurations. Indeed, looking at the plot in Figure 2, we note that the break is rather a subtle one, with the estimated slopes differing by approximately 0.37. In summary, for this particular dataset we note that there appears to be evidence of a breakpoint, although the sample size required to detect the breakpoint is not reached until we truncate at $K = 8$, just before additional modeling considerations such as incompleteness must be accounted for.

7 Theoretical Properties

This section deals with the large-sample properties of the proposed procedure. In order to make the proofs more accessible, we first focus on the case when B is known, with no background contamination ($b_i = 0$ for all i) and all A_i are assumed to be identical. The more general cases will be briefly discussed later.

If it is assumed that $A_i = A > 0$ for all $i = 1, \dots, n$, then Y_1, \dots, Y_n constitute an iid sample from the previously stated statistical model. Denote the density of Y_1 by

$$\begin{aligned} f(y; \boldsymbol{\theta}) &= \int_{\tau_1}^{\infty} \frac{e^{-As}(As)^y}{y!} f_B(s; \boldsymbol{\beta}, \boldsymbol{\tau}) ds \\ &= \sum_{j=1}^B \left(\frac{\tau_{j-1}}{\tau_j} \right)^{\beta_{j-1}} \frac{\beta_j (A\tau_j)^{\beta_j}}{y!} \{ \Gamma(y - \beta_j, A\tau_j) - \Gamma(y - \beta_j, A\tau_{j+1}) \} \\ &= \sum_{j=1}^B \left(\frac{\tau_{j-1}}{\tau_j} \right)^{\beta_{j-1}} \frac{\beta_j (A\tau_j)^{\beta_j}}{y!} \int_{A\tau_j}^{A\tau_{j+1}} t^{y-\beta_j-1} e^{-t} dt. \end{aligned}$$

The parameter space is defined as $\Theta = \{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})^T \in \mathbb{R}_+^{2B} : \beta_j \neq \beta_{j+1}, \tau_j < \tau_{j+1}, j = 1, \dots, B-1\}$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\tau}_0)^T \in \Theta$ denote the true parameter value. Notice that Θ is not compact and that the value of the likelihood does not converge to zero if the parameter approaches the boundary of Θ . Therefore standard arguments such as the ones based on Wald (1949) do not apply directly in order to establish strong consistency of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$. Instead a compactification device is applied to subsequently use the results of Kiefer and Wolfowitz (1956). This leads to the following result.

Theorem 1. *Suppose B is known and $A_i = A > 0$ for all $i = 1, \dots, n$. Then, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is strongly consistent for $\boldsymbol{\theta}_0$, that is, $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$ with probability one as $n \rightarrow \infty$.*

The proof of Theorem 1 is provided in Appendix A. To weaken the restriction of identical A_i , observe that this condition is mainly applied to allow the use of the strong law of large numbers for iid random variables, as required for the direct use of the results in Wald (1949) and Kiefer and Wolfowitz (1956). Since the arguments used to prove Theorem 1 are still valid if only the assumption $A_i > 0$ is made, Kolmogorov's version of the strong law of large numbers can be applied to adapt their proof to the present case, imposing additional assumptions such as the Kolmogorov criterion

$$\sum_{i=1}^{\infty} \frac{\text{Var}(Y_i)}{i^2} < \infty$$

or conditions ensuring the validity of Kolmogorov’s three-series theorem. Then, the result of Theorem 1 holds also in this more general setting. The case for non-zero b_i ’s can also be dealt with similarly, but with long and tedious algebra.

In the theory developed above, the number of pieces, B , in the broken-Pareto model is assumed to be known. The case of unknown B is, however, substantially more difficult. In fact, results in simpler settings such as the traditional “change in mean” scenario, in which segments of independent observations differ only by their levels, strong distributional assumptions become necessary to show consistency of an estimator for B . These typically require normality of the observations so that sharp tail estimates of the supremum of certain Gaussian processes are available; e.g., see Yao (1988). These techniques have also been exploited in Aue and Lee (2011) for image segmentation purposes. However, in the current context of the more complex broken-Pareto model, these arguments are not applicable and in fact it seems infeasible to derive theoretical properties under a set of practically relevant assumptions.

8 Concluding Remarks

We provide a coherent statistical procedure for selecting the number and orientation of “pieces” in an assumed piecewise linear $\log N - \log S$ relationship. Our framework allows astrophysicists to use a principled approach to reliably select the model order B , and for parameter estimation via maximum likelihood estimation in a numerically challenging context. To the best of our knowledge, this is the first statistically rigorous procedure developed for solving this important scientific problem. *R*-codes implementing the proposed procedure can be obtained from the authors.

A Technical Details

To prove Theorem 1, the five assumptions made in Section 2 of Kiefer and Wolfowitz (1956) need to be verified. This is done in the following.

Assumption 1. *It is required that $f(y; \theta)$ is a density with respect to a σ -finite measure μ on a Euclidean space of which y is the generic point.*

Proof. This condition is satisfied since the underlying distribution is discrete. □

Define a metric on the space Θ by setting

$$\delta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{j=1}^B |\arctan \beta_{1,j} - \arctan \beta_{2,j}| + \sum_{j=1}^B |\arctan \tau_{1,j} - \arctan \tau_{2,j}|.$$

Following Kiefer and Wolfowitz (1956), the parameter space is compactified by defining $\bar{\Theta}$ to be the completion of Θ by adding all the limits of its Cauchy sequences in the sense of the above metric. Unless otherwise mentioned, all limits involving $\boldsymbol{\theta}$ are understood to be with respect to δ . The Euclidean norm is denoted by $|\cdot|_E$. To verify the next assumption of Kiefer and Wolfowitz (1956), two auxiliary lemmas are introduced.

Lemma 1. *For sufficiently large β_j and a fixed $y \in \mathbb{N}_0$, we have*

$$\beta_j(A\tau_j)^{\beta_j} \int_{A\tau_j}^{A\tau_j+1} t^{y-\beta_j-1} e^{-t} dt < 2(A\tau_j)^y e^{-A\tau_j},$$

where $\boldsymbol{\tau} \in \Theta$.

Proof. Note that $\beta_j(A\tau_j)^{\beta_j} \int_{A\tau_j}^{A\tau_j+1} t^{y-\beta_j-1} e^{-t} dt \leq \beta_j(A\tau_j)^{\beta_j} \Gamma(y - \beta_j, A\tau_j)$. Thus, by Theorem 2.2 of Borwein and Chan (2009), for a sufficiently large β_j and a fixed $y \in \mathbb{N}_0$,

$$\Gamma(y - \beta_j, A\tau_j) \leq \frac{-(A\tau_j)^{y-\beta_j} e^{-A\tau_j}}{y - \beta_j}$$

and consequently $\beta_j(A\tau_j)^{\beta_j} \Gamma(y - \beta_j, A\tau_j) < 2(A\tau_j)^y e^{-A\tau_j}$. \square

Lemma 2. *If $|\boldsymbol{\tau}|_E < \infty$, $\lim_{|\boldsymbol{\beta}|_E \rightarrow \infty} f(y; \boldsymbol{\theta})$ exists.*

Proof. Note that if $|\boldsymbol{\beta}|_E \rightarrow \infty$, there exists a j such that $\beta_j \rightarrow \infty$. We focus on that one particular j . Let $g_j(y; \boldsymbol{\theta}) = \beta_j(A\tau_j)^{\beta_j} \int_{A\tau_j}^{A\tau_j+1} t^{y-\beta_j-1} e^{-t} dt$. In order to show that the limit of f exists, we only have to show that the limit of g exists (since it generalizes to any j with $\beta_j \rightarrow \infty$). Note that, instead of considering g_j , we look at $h_j(y; \boldsymbol{\theta}) = \log g_j(y; \boldsymbol{\theta})$. We define $h_{j,1}(y; \boldsymbol{\theta}) = \log\{\beta_j(A\tau_j)^{\beta_j}\}$ and $h_{j,2}(y; \boldsymbol{\theta}) = \log \int_{A\tau_j}^{A\tau_j+1} t^{y-\beta_j-1} e^{-t} dt$. Then we have

$$\begin{aligned} \frac{\partial h_j(y; \boldsymbol{\theta})}{\partial \beta_j} &= \frac{\partial h_{j,1}(y; \boldsymbol{\theta})}{\partial \beta_j} + \frac{\partial h_{j,2}(y; \boldsymbol{\theta})}{\partial \beta_j} \\ &= \left\{ \frac{1}{\beta_j} + \log(A\tau_j) \right\} + \left\{ -\frac{\int_{A\tau_j}^{A\tau_j+1} t^{y-\beta_j-1} e^{-t} (\log t) dt}{\int_{A\tau_j}^{A\tau_j+1} t^{y-\beta_j-1} e^{-t} dt} \right\} \\ \frac{\partial^2 h_{j,1}(y; \boldsymbol{\theta})}{\partial \beta_j^2} &= -\frac{1}{\beta_j^2} \end{aligned}$$

$$\frac{\partial^2 h_{j,2}(y; \boldsymbol{\theta})}{\partial \beta_j^2} = \frac{\int_{A\tau_j}^{A\tau_{j+1}} t^{y-\beta_j-1} e^{-t} (\log t)^2 dt}{\int_{A\tau_j}^{A\tau_{j+1}} t^{y-\beta_j-1} e^{-t} dt} - \left\{ \frac{\int_{A\tau_j}^{A\tau_{j+1}} t^{y-\beta_j-1} e^{-t} (\log t) dt}{\int_{A\tau_j}^{A\tau_{j+1}} t^{y-\beta_j-1} e^{-t} dt} \right\}^2$$

Note that $\partial^2 h_{j,2}(y; \boldsymbol{\theta}) / \partial \beta_j^2 = \text{Var}(\log(T))$, where T is a random variable with density

$$r(t) = \frac{t^{y-\beta_j-1} e^{-t}}{\int_{A\tau_j}^{A\tau_{j+1}} s^{y-\beta_j-1} e^{-s} ds}, \quad A\tau_j < t < A\tau_{j+1}.$$

Thus $\partial^2 h_{j,2}(y; \boldsymbol{\theta}) / \partial \beta_j^2 \geq 0$ and $\partial h_{j,2}(y; \boldsymbol{\theta}) / \partial \beta_j$ is increasing with respect to β_j . It follows then that $\partial h_{j,2}(y; \boldsymbol{\theta}) / \partial \beta_j$ is bounded from above since h_j is bounded from above by Lemma 1. Consequently, $\lim_{\beta_j \rightarrow \infty} \partial h_{j,2}(y; \boldsymbol{\theta}) / \partial \beta_j$ exists. \square

Assumption 2 (Continuity Assumption). *It is possible to extend the definition of $f(y; \boldsymbol{\theta})$ so that the range of $\boldsymbol{\theta}$ will be $\bar{\Theta}$ and so that, for any $\{\boldsymbol{\theta}_i\}$ and $\boldsymbol{\theta}^*$ in $\bar{\Theta}$, $\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}^*$ implies*

$$f(y; \boldsymbol{\theta}) \rightarrow f(y; \boldsymbol{\theta}^*)$$

except perhaps on a set of y whose probability is 0 according to the probability density $f(y; \boldsymbol{\theta}_0)$. (The exceptional y -set may depend on $\boldsymbol{\theta}^$ and $f(y; \boldsymbol{\theta}^*)$ need not be a probability density function.)*

Proof. First, $f(y; \boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta} \in \Theta$ and thus f automatically fulfills the above continuity requirement for $\boldsymbol{\theta} \in \Theta$. Define $\partial\Theta = \bar{\Theta} \setminus \Theta$. Now, we will show that we can define $f(y; \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^* \in \partial\Theta$, as $\lim_{\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*} f(y; \boldsymbol{\theta})$. It is thus only required to show the existence of this limit. Notice that $\lim_{\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*} f(y; \boldsymbol{\theta})$ exists for boundary points $\boldsymbol{\theta} \in \partial\Theta$ with $|\boldsymbol{\theta}|_E \neq \infty$. The remaining case $|\boldsymbol{\theta}|_E = \infty$ can be separated into three sub-cases: (i) $|\boldsymbol{\beta}|_E = \infty$ and $|\boldsymbol{\tau}|_E < \infty$, (ii) $|\boldsymbol{\beta}|_E < \infty$ and $|\boldsymbol{\tau}|_E = \infty$, and (iii) $|\boldsymbol{\beta}|_E = \infty$ and $|\boldsymbol{\tau}|_E = \infty$.

1. Suppose $|\boldsymbol{\beta}|_E = \infty$ and $|\boldsymbol{\tau}|_E < \infty$. From Lemma 2, $\lim_{|\boldsymbol{\beta}|_E \rightarrow \infty} f(y; \boldsymbol{\theta})$ exists.
2. Suppose $|\boldsymbol{\beta}|_E < \infty$ and $|\boldsymbol{\tau}|_E = \infty$. This implies that there exists at least one j such that $\tau_j = \infty$. Here, we have

$$0 \leq a^{\beta_j} \int_{Aa}^{Ab} t^{y-\beta_j-1} e^{-t} dt \leq a^{\beta_j} \int_{Aa}^{\infty} t^{y-\beta_j-1} e^{-t} dt,$$

where $0 < a < b$. Taking the limit on the right-hand side, using the l'Hospital rule, it follows

that

$$\lim_{a \rightarrow \infty} \frac{\int_{Aa}^{\infty} t^{y-\beta_j-1} e^{-t} dt}{a^{-\beta_j}} = \lim_{a \rightarrow \infty} \frac{A^{y-\beta_j-1} \tau_j^y e^{-Aa}}{\beta_j} = 0.$$

Since $0 \leq (c/a)^{\beta_j-1} \leq 1$ for all $0 < c < a$, $\lim_{|\tau|_E \rightarrow \infty} f(y; \theta)$ exists.

3. $|\beta|_E = \infty$ and $|\tau|_E = \infty$. The existence of $\lim_{|\theta|_E \rightarrow \infty} f(y; \theta)$ is basically implied by Lemma 1.

The proof is complete. \square

Assumption 3. For any θ in $\bar{\Theta}$ and any $\rho > 0$, $w(y; \theta, \rho)$ is a measurable function of y , where

$$w(y; \theta, \rho) = \sup f(y; \theta'),$$

the supremum being taken over all θ' in $\bar{\Theta}$ for which $\delta(\theta, \theta') < \rho$.

Proof. The statement is implied by the continuity of $f(y; \theta)$ with respect to $\theta \in \bar{\Theta}$. \square

Assumption 4 (Identifiability Assumption). If θ in $\bar{\Theta}$ is different from θ_0 , then, for at least one x ,

$$\int_{-\infty}^x f(y|\theta) d\mu \neq \int_{-\infty}^x f(y|\theta_0) d\mu,$$

the integral being over those y all of whose components \leq the corresponding of x .

Proof. In the present case, μ is the counting measure and thus, for all $\theta \in \bar{\Theta}$, if $f(y|\theta) \neq f(y|\theta_0)$ for at least one $y \in \mathbb{N}_0$, it fulfills the above assumption. This is obviously true for $\theta \in \Theta$. Since $\theta_0 \in \Theta$, it is also easy to see that the above is true for $\theta \in \bar{\Theta}$. \square

Assumption 5 (Integrability Assumption). For any θ in $\bar{\Theta}$ we have

$$\lim_{\rho \downarrow 0} \mathbb{E} \left[\log \frac{w(Y; \theta, \rho)}{f(Y; \theta_0)} \right]^+ < \infty,$$

where w is defined in Assumption 3.

Proof. Since $f(y; \theta)$ is continuous and bounded over $\bar{\Theta}$, $\log w(y; \theta, \rho)$ is bounded from above. Now, we want to show that $\mathbb{E} |\log f(Y; \theta_0)| < \infty$. Since $f(y; \theta_0)$ is bounded from above, we only need

$\mathbb{E}\{\log(f(Y; \boldsymbol{\theta}_0))\} > -\infty$, which can be shown as follows. Note that, for any $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}
\mathbb{E}[\log\{f(Y; \boldsymbol{\theta})\}] &= \mathbb{E} \left[\log \left\{ \sum_{j=1}^B \left(\frac{\tau_{j-1}}{\tau_j} \right)^{\beta_{j-1}} \frac{\beta_j (A\tau_j)^{\beta_j}}{Y!} \int_{A\tau_j}^{A\tau_{j+1}} t^{Y-\beta_j-1} e^{-t} dt \right\} \right] \\
&\geq \sum_{j=1}^B \left[\log \left\{ \left(\frac{\tau_{j-1}}{\tau_j} \right)^{\beta_{j-1}} \beta_j (A\tau_j)^{\beta_j} \right\} + \mathbb{E} \left\{ \log \left(\frac{\int_{A\tau_j}^{A\tau_{j+1}} t^{Y-\beta_j-1} e^{-t} dt}{Y!} \right) \right\} \right] \\
&\geq \sum_{j=1}^B \log \left\{ \left(\frac{\tau_{j-1}}{\tau_j} \right)^{\beta_{j-1}} \beta_j (A\tau_j)^{\beta_j} \right\} + \sum_{j=1}^B \mathbb{E} \left\{ \log \left(\frac{\int_{A\tau_j}^{\infty} t^{Y-\beta_j-1} e^{-t} dt}{Y!} \right) \right\} \\
&= \sum_{j=1}^B \log \left\{ \left(\frac{\tau_{j-1}}{\tau_j} \right)^{\beta_{j-1}} \beta_j (A\tau_j)^{\beta_j} \right\} + \sum_{j=1}^B \mathbb{E} \left[\log \left\{ \frac{\Gamma(Y - \beta_j, A\tau_j)}{\Gamma(Y + 1)} \right\} \right]
\end{aligned}$$

Here,

$$\frac{\Gamma(Y - \beta_j, A\tau_j)}{\Gamma(Y + 1)} = \frac{\Gamma(Y - \beta_j, A\tau_j)}{\Gamma(Y - \beta_j)} \frac{\Gamma(Y - \beta_j)}{\Gamma(Y + 1)} = Q(Y - \beta_j, A\tau_j) \frac{\Gamma(Y - \beta_j)}{\Gamma(Y + 1)},$$

where Q is the regularized incomplete gamma function. Now, we state the asymptotic expansions of the regularized incomplete gamma function and the ratio of two gamma functions: When $a \rightarrow \infty$,

$$\begin{aligned}
Q(a, z) &\propto 1 - \frac{a^{-a-1/2} e^{a-z} z^a}{\sqrt{2\pi}} \left\{ 1 + O\left(\frac{1}{a}\right) \right\}, \\
\frac{\Gamma(a+b)}{\Gamma(a+c)} &\propto a^{b-c} \left\{ 1 + O\left(\frac{1}{a}\right) \right\}.
\end{aligned}$$

Applying these asymptotic expansions for large y ,

$$\frac{\Gamma(y - \beta_j, A\tau_j)}{\Gamma(y + 1)} \propto y^{-\beta_j-1} \left\{ 1 + O\left(\frac{1}{y}\right) \right\}.$$

Thus, in order to bound $\mathbb{E}[\log\{\Gamma(Y - \beta_j, A\tau_j)/\Gamma(Y + 1)\}]$, we only have to bound $\mathbb{E}\{\log(Y)1_{\{Y \geq M\}}\}$ away from ∞ for sufficiently large M . Here, we define $\log 0 \times 0 = 0$. Now, we only have to consider the boundedness of $\sum_{y=M}^{\infty} \log(y)/y^{\beta_j+1}$ for $j = 1, \dots, B$. It is bounded whenever $\beta_j > 0$, which is fulfilled by any $\boldsymbol{\theta} \in \Theta$. Thus, $\mathbb{E}\{\log(f(Y; \boldsymbol{\theta}_0))\} > -\infty$ since $\boldsymbol{\theta}_0 \in \Theta$. \square

The statement of Theorem 1 follows now from Section 2 of Kiefer and Wolfowitz (1956).

References

- Aue, A. and Lee, T. C. M. (2011) On image segmentation using information theoretic criteria. *The Annals of Statistics*, **39**, 2912–2935.
- Baines, P. D. (2010) *Statistics, Science and Statistical Science: Modeling, Inference and Computation with Applications to the Physical Sciences*. Ph.D. thesis.
- Baines, P. D., Meng, X.-L. and Xie, X. (2012) The interwoven EM algorithm. Submitted for publication.
- Borwein, J. M. and Chan, O.-Y. (2009) Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities & Applications*, **12**, 115–121.
- Cappelluti, N., Hasinger, G., Brusa, M., Comastri, A., Zamorani, G., Bhringer, H., Brunner, H., Civano, F., Finoguenov, A., Fiore, F., Gilli, R., Griffiths, R. E., Mainieri, V., Matute, I., Miyaji, T. and Silverman, J. (2007) The XMM-Newton Wide-Field Survey in the COSMOS Field. II. X-ray data and the log N–log S relations. *The Astrophysical Journal Supplement Series*, **172**, 341.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- Friel, N. and Pettitt, A. N. (2008) Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society, Series B*, **70**, 589–607.
- Guetta, D., Granot, J. and Begelman, M. C. (2005) Constraining the structure of gamma-ray burst jets through the log N–log S distribution. *The Astrophysical Journal*, **622**, 482–491.
- Hickox, R. C. and Markevitch, M. (2007) Can Chandra resolve the remaining cosmic X-ray background? *The Astrophysical Journal*, **671**, 1523–1530.
- Jordán, A., Côté, P., Ferrarese, L., Blakeslee, J. P., Mei, S., Merritt, D., Milosavljević, M., Peng, E. W., Tonry, J. L. and West, M. J. (2004) The ACS Virgo Cluster Survey. III. Chandra and Hubble space telescope observations of low-mass X-ray binaries and globular clusters in M87. *The Astrophysical Journal*, **613**, 279.
- Kenter, A. T. and Murray, S. S. (2003) A new technique for determining the number of X-ray sources per flux density interval. *The Astrophysical Journal*, **584**, 1016–1020.

- Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, **27**, 887–906.
- Kitayama, T., Sasaki, S. and Suto, Y. (1998) Cosmological implications of number counts of clusters of galaxies: log N–log S in X-ray and submm bands. *Publications of the Astronomical Society of Japan*, **50**, 1–11.
- Mateos, S., Warwick, R. S., Carrera, F. J., Stewart, G. C., Ebrero, J., Della Ceca, R., Caccianiga, A., Gilli, R., Page, M. J., Treister, E., Tedds, J. A., Watson, M. G., Lamer, G., Saxton, R. D., Brunner, H. and Page, C. G. (2008) High precision X-ray log N–log S distributions: Implications for the obscured AGN population. *Astronomy & Astrophysics*, **492**, 51–69.
- Mathiesen, B. and Evrard, A. E. (1998) Constraints on Ω_0 and cluster evolution using the ROSAT log N–log S relation. *Monthly Notices of the Royal Astronomical Society*, **295**, 769–780.
- Moretti, A., Campana, S., Lazzati, D. and Tagliaferri, G. (2003) The resolved fraction of the cosmic X-ray background. *The Astrophysical Journal*, **588**, 696–703.
- Trudolyubov, S. P., Borozdin, K. N., Priedhorsky, W. C., Mason, K. O. and Cordova, F. A. (2002) On the X-ray source luminosity distributions in the bulge and disk of M31: First results from the XMM-Newton Survey. *The Astrophysical Journal Letters*, **571**, 17–21.
- Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, **20**, 595–601.
- Yao, Y. C. (1988) Estimating the number of change-points via Schwarz’ criterion. *Statistics & Probability Letters*, **6**, 181–189.
- Yu, Y. and Meng, X.-L. (2011) To center or not to center: That is not the question — An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, **20**, 531–570.